# Article Watch: September 2019

*Clive A. Slaughter*

This column highlights recently published articles that are of interest to the readership of this publication. We encourage ABRF members to forward information on articles they feel are important and useful to Clive Slaughter, MCG-UGA Medical Partnership, 1425 Prince Avenue, Athens, GA 30606, USA. Tel: (706) 713-2216; Fax: (706) 713-2221; E-mail: cslaught@uga.edu, or to any member of the editorial board. Article summaries reflect the reviewer's opinions and not necessarily those of the Association.

## NUCLEIC ACID SEQUENCING AND GENOTYPING

**Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res* 2019;29:954-960.**

Breitwieser *et al.* have conducted a systematic search of the bacterial and archaeal sections of the NCBI RefSeq genome database for contamination by human sequences. Human DNA is almost always present in the environment of sequencing laboratories, but validation procedures are expected to identify human contamination of nonhuman data sets. However, the current assembly of the human genome is still incomplete, particularly in missing some repeat-rich regions. Variation among the many copies of these repeats introduces sequences of human origin that do not match precisely to the human reference genome. They are therefore difficult to detect as contaminants. The authors therefore employ profile hidden Markov models of human repeats in their search for human contamination. They identify 2250 microbial genomes that are contaminated by human sequences. A further problem arises when contigs containing such human sequences are annotated as protein-coding genes. The cognate protein sequences may then be incorporated into other databases and the spurious proteins used for future annotation. Spurious protein entries numbering 3437 are found to be present in the NCBI nonredundant protein database and TrEMBL protein database (*https://www.uniprot.org/statistics/TrEMBL*) that purport to represent protein families spanning prokaryotic and some eukaryotic genomes. Such erroneous information may lead to incorrect biologic conclusions. For example, when metagenomic sequencing is used for the diagnosis of infection, if a microbe contains unrecognized fragments of the human genome, the presence of these fragments in samples of human DNA might be concluded to indicate infection with that microbe. Similarly, the presence of fragments of the wrong species appearing in a genome may falsely be concluded to indicate horizontal gene transfer. Because most contamination appears in small, low-coverage contigs in draft genomes, the authors recommend that small contigs be filtered from draft genome assemblies to avoid future instances of such contamination.

doi: 10.7171/jbt.19-3003-003

## GLYCANS

**Chang MM, Gaidukov L, Jung G, et al. Small-molecule control of antibody N-glycosylation in engineered mammalian cells. Nat Chem Biol 2019; 15:730-736.**

Control of glycosylation levels and of glycan structure in mAb Fc regions is an enduring challenge in the manufacture of biologic pharmaceuticals. Chang *et al.* seek to exercise precise quantitative control of the activity of enzymes in the glycosylation pathway in order to test new antibody therapeutics or improve existing ones. They knock out 2 glycosyltransferase genes in Chinese hamster ovary (CHO) cells, α-1,6-fucosyltransferase (*FUT8*) and β-1,4-galactosyltransferase (*β4GALT1*), and replace them with synthetic glycosyltransferase genes under constitutive or inducible promoters. This enables simultaneous and independent induction of fucosylation (0–95%) and galactosylation (0–87%) at tunable levels by the inducers doxycycline and abscisic acid. These manipulations facilitate investigation of the effects of *N*-glycan profiles on effector functions of antibodies.

**Engle DD, Tiriac H, Rivera KD, et al. The glycan CA19-9 promotes pancreatitis and pancreatic cancer in mice. Science 2019;364:1156-1162.**

The glycan carbohydrate antigen 19-9 (CA19-9, otherwise known as sialyl-Lewis$^{a}$) has long been recognized as a biomarker for pancreatic disease. It is found at increased levels in the serum of 10–30% of patients with pancreatitis and in 75% of patients with pancreatic cancer. CA19-9 is a glycan moiety found on proteins, lipids, and other glycans. Its synthesis depends on fucosyltransferase 3 (FUT3) and β-1,3-galactosyltransferase 5 (β3GALT5). Mice lack FUT3 and therefore do not make CA19-9. As part of an effort to discover candidate biomarkers for pancreatic cancer, Engle *et al.* created a mouse that inducibly expresses human FUT3 and β3GALT5, and can therefore make CA19-9. These mice are found to modify the extracellular matrix protein fibrilin-3 with CA19-9 moieties. This results in rapid onset of pancreatitis associated with activation of epidermal growth factor receptor (EGFR) signaling. EGFR activation results from increased interaction between the modified fibrillin and EGFR. The phenotype can be reversed by shutting down CA19-9 synthesis or by blocking CA19-9. The authors also demonstrate that EGFR activation driven by CA19-9 stimulates *Kras*$^{G12D}$-driven tumorigenesis in these mice. This study indicates a specific functional role for a glycan moiety in human disease and suggests CA19-9 as a potential therapeutic target.

## MACROMOLECULAR SYNTHESIS AND SYNTHETIC BIOLOGY

**Fredens J, Wang K, de la Torre D, et al. Total synthesis of Escherichia coli with a recoded genome. Nature 2019;569:514-518.**

Fredens *et al.* have created Syn61, a variant of *Escherichia coli*, a bacterium with a 4-Mb genome, in which the entire genome of the prototypical *E. coli* strain MDS42 is replaced with synthetic DNA. Their work explores the limits of changes in codon usage. The 20 amino acids, plus "'start"' and "'stop"' signals, are normally encoded by 64 codons. Syn61 uses only 61 codons; 2 of the 6 alternative codons for serine and the stop codon TAG (the amber stop codon, of which there are only 321 instances in the genome) are completely replaced by synonymous codons. This required recoding 18,214 codons altogether. One previously essential transfer RNA may be deleted or repurposed in Syn61. The synthetic strain is viable, although doubling times are somewhat longer than those of MDS42. This work represents a new milestone in synthetic biology.

**Zhao EM, Suek N, Wilson MZ, et al. Light-based control of metabolic flux through assembly of synthetic organelles. Nat Chem Biol 2019;15:589-597.**

The goal of metabolic engineering, to maximize the formation of a desired product, is usually achieved by overexpressing the enzymes that synthesize that product. However, yield of product may be limited if constitutive diversion of flux toward product synthesis starves pathways essential for the cell's normal functions. Dynamic regulation of metabolic flux, plus colocalization of the relevant enzymes to increase the efficiency of product formation, mitigates such barriers. Zhao *et al.* exploit optogenetic control in yeast to assemble or disassemble active enzyme clusters for these purposes. They demonstrate that light-controlled clustering of 2 enzymes in the deoxyviolacein pathway, plus careful control of the enzymes' synthesis rate, enhances metabolic flux in a light-switchable manner and thereby increases the yield of product by 6-fold. Clustering into metabolic organelles is also generally advantageous because it diminishes the intracellular concentration of intermediary metabolites in the synthesis pathway. Such intermediates may be cytotoxic or otherwise reduce yield by siphoning off flux from the desired pathway.

## MASS SPECTROMETRY

**Gessulat S, Schmidt T, Zolg DP, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods 2019;16:509-518.**

**Tiwary S, Levy R, Gutenbrunner P, et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat Methods 2019;16:519-525.**

**Kirik U, Refsgaard JC, Jensen LJ. Improving peptide-spectrum matching by fragmentation prediction using hidden Markov models. J Proteome Res 2019;18:2385-2396.**

In proteomics, protein identification is most often performed by tryptic digestion, separation of the resulting peptides by liquid chromatography, gas-phase peptide fragmentation, and database matching of the observed fragment ions with the mass values computed using rules governing the fragmentation process. This workflow generally ignores the relative intensities of the fragment ion signals, which are determined by the probability of cleavage at the different peptide bonds comprising the precursor peptide. These intensities nevertheless contain much information of potential value for peptide identification. For the purpose of exploiting this largely unutilized source of information, 3 groups now announce the results of studies to predict fragment ion intensities using models constructed by training on large data repositories of observed peptide fragmentation patterns. All 3 groups observe high predictive performance. They note that their models accommodate peptides of any length and enzymic cleavage process as well as multiple fragmentation techniques. Gessulat *et al.* additionally train their network to predict retention times on reverse-phase chromatography. This group has integrated its neural network into the ProteomicsDB (*https://www.proteomicsdb.org/prosit*) to allow users to rescore their search results. Tiwary *et al.* plan to integrate their neural network into MaxQuant. Kirik *et al.* use a hidden Markov model for prediction of fragment intensities. They advocate prefiltering product ions to eliminate fragments likely to be of low intensity in order to improve statistical performance of database searching. It is anticipated that these prediction methods will become especially useful in future when they are applied in circumstances in which search database sizes are extremely large. This occurs, for example, in searches involving post-translational modifications, unknown proteolytic cleavages, and metaproteomics.

## METABOLOMICS

**Neumann EK, Ellis JF, Triplett AE, Rubakhin SS, Sweedler JV. Lipid analysis of 30,000 individual rodent cerebellar cells using high-resolution mass spectrometry. Anal Chem 2019;91:7871-7878.**

Neumann *et al.* undertake large-scale, single-cell profiling of lipids in the brain to assess heterogeneity among cells. Lipids represent a hitherto-understudied class of cellular constituents in the single-cell arena. The authors deploy microscopy-guided matrix-assisted laser desorption–ionization (MALDI) mass spectrometry for single-cell profiling and determine the lipid composition of 30,000 individual cells of the rat cerebellum using this platform. Sampling from single cells is ensured by proteolytic dissociation of the tissue and diffuse deposition of the resulting cell suspension on MALDI targets. The small quantities of lipid available from single cells preclude tandem mass spectrometry for feature assignment. However, using the high resolution and mass accuracy provided by ion cyclotron resonance mass spectrometry, the authors putatively identify 500 lipid spectral features without fragmentation based on comparison with information derived from liquid chromatography–tandem mass spectrometry (LC/MS/MS) databases and mass spectral data sets of brain lipids. Clustering analysis enabled 101 distinct lipid assemblages to be distinguished. The numbers of cells belonging to each assemblage pattern varied from 882 cells for the most common to 31 cells for the rarest. Fewer than 30 canonical cell types have been described in the cerebellum. This suggests that the analysis is detecting functional or cell-state heterogeneity of interest for future detailed exploration.

**Fu X, Deja S, Kucejova B, Duarte JAG, McDonald JG, Burgess SC. Targeted determination of tissue energy status by LC-MS/MS. Anal Chem 2019;91: 5881-5887.**

The energy status of cells may, in principle, be characterized in terms of the abundance of a set of intracellular nucleotides and acyl-CoAs. However, no single method has hitherto been described for simultaneous measurement of the requisite metabolites. Fu *et al.* here describe such a method. The procedure is based on ion-pairing reverse-phase liquid chromatography and online electrospray ionization mass spectrometry. The authors measure the adenine nucleotides AMP, ADP, and ATP; the pyridine dinucleotides $NAD^+$ and NADPH; and short-chain acyl-CoAs (acetyl, malonyl, succinyl, and propionyl). They demonstrate the utility of the method in studies of mouse liver tissue following various periods of ischemia. Hypoxia-induced changes are detected in the tissue following blood loss, indicating the need for caution in interpretation of metabolomic data acquired when multiple organs are sampled from an experimental animal.

## FUNCTIONAL GENOMICS AND PROTEOMICS

**Yizhak K, Aguet F, Kim J, et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. Science 2019;364: eaaw0726.**

Yizhak *et al.* conduct an extensive survey of normal adult tissues for somatic variants occurring in macroscopic clones. Rather than focusing on a small number of preselected genes, they wish to sample genes more widely. For this purpose, they developed an analysis pipeline to detect somatic mutations from RNA sequencing (RNA-Seq) data and apply the pipeline to data collected for 29 normal tissues from 488 individuals in the Genotype-Tissue Expression (GTEx) project. They employ training data sets in which mutations in RNA and DNA can be compared. They establish that RNA-Seq provides a sensitive and precise method for mutation detection in genes that are expressed strongly enough to give adequate sequencing depth. The data are filtered firstly to remove alignment errors by using 2 different RNA aligners, secondly to remove sequencing errors using a site-specific error model, and thirdly to remove RNA editing sites using databases of known editing sites. The incidence of somatic mutations is found to be very high. Mutations are detected in almost all individuals and tissues studied. Sun-exposed skin, esophagus mucosa, and lung show the largest number of mutations, suggesting the involvement of environmental exposure.

Numbers of mutations also depended upon the age of the individual and on the rate of cellular turnover in the tissue. Some of the mutations are in cancer-associated genes, although no cancer was present. Yet evidence for a proliferative advantage in the mutated clones is present in some instances. Because the methodology used by the authors favors discovery of somatic mutation in larger clones, the reservoir of mutations in smaller clones is expected to be even more extensive. This foundational study opens numerous avenues for future investigation of the process of cellular transformation and the clinical detection of cancer. It provides new methodology for pursuit of these avenues.

## PROTEOMICS

**Zecha J, Satpathy S, Kanashova T, et al. TMT labeling for the masses: a robust and cost-efficient, in-solution labeling approach. Mol Cell Proteomics 2019;18:1468-1478.**

Tandem mass tags (TMTs) are widely used for multiplexed quantification of peptides in proteome digests. The TMT reagents (amine-reactive NHS esters) are costly. In the interest of cost saving, Zecha *et al.* systematically investigate labeling reaction parameters to establish how little of the TMT reagents may be provided without sacrificing labeling efficiency. The authors test different TMT and peptide concentrations, quantities, and ratios. They find that when the reaction volume is reduced to keep peptide and TMT concentrations favorably high (3.4 μg/μL and 2 g/L, respectively), TMT-to-peptide ratios as low as 1:1 (wt:wt) can be used without diminution of labeling efficiency. The authors' protocol uses 8 times less TMT reagent than is recommended by the manufacturer. The authors comment that the same principles may also apply to other NHS ester reactions, such as iTRAQ labeling, biotinylation, and cross-linking.

## CELL BIOLOGY AND TISSUE ENGINEERING

**Manfrin A, Tabata Y, Paquet ER, et al. Engineered signaling centers for the spatially controlled patterning of human pluripotent stem cells. Nat Methods 2019;16:640-648.**

In the developmental process of gastrulation, a uniform ball of embryonic stem cells forms the primary germ layers (ectoderm, mesoderm, *etc.*). These cell types are distributed in concentric layers that subsequently form asymmetric patterns to establish the axes of the body (dorsal-ventral, anterior-posterior). Spatial distributions are established by the cells' responses to gradients of morphogens synthesized in signaling centers. Culture conditions that replicate the localized morphogen signaling required for correct spatial patterning have not hitherto been achieved. Manfrin *et al.* here report a microfluidic system for emulating signaling centers in stem cell culture by spatially and temporally controlling morphogen diffusion from a localized source. Using bone morphogenetic protein 4 (BMP4) as the morphogen, the authors demonstrate that human pluripotent stem cell colonies reproducibly form asymmetric structures in response to the morphogen concentration gradients. The authors further enhance spatial control of patterning by introducing gradients of a BMP4 inhibitor. Their experimental system is hoped to help answer questions related to the effects of cell density, temporal variation in signaling, and the production of paracrine factors in germ layer patterning.

**Xiang C, Du Y, Meng G, et al. Long-term functional maintenance of primary human hepatocytes in vitro. Science 2019;364:399-402.**

Terminally differentiated cells tend to lose characteristic functions and gene expression signatures in culture for want of the signals and microenvironmental cues that stabilize their differentiated state *in vivo*. Xiang *et al.* have identified 5 small molecules that modulate known cellular targets that, when supplied together, stabilize the differentiated state of primary human hepatocytes for several weeks *in vitro*. Treated cells support the complete 4-wk infection cycle of hepatitis B virus and therefore are suitable for testing new antiviral strategies.

## IMAGING

**Taylor RW, Mahmoodabadi RG, Rauschenberger V, Giessl A, Schambony A, Sandoghdar V. Interferometric scattering microscopy reveals microsecond nanoscopic protein motion on a live cell membrane. Nat Photonics 2019;13:480-487.**

Methods based on fluorescence microscopy continue to evolve for tracking the motions of proteins in live cells. However, fluorescence methods are ultimately limited in the temporal resolution and duration of imaging they can achieve. The limitation derives from the physical characteristics of fluorescence emission. The present paper describes recent progress by a group exploring the use of Rayleigh scattering in place of fluorescence emission to track the movement of particles in cells. Their technique of interferometric scattering microscopy is sensitive, but the chief difficulty with its use on live cells derives from a strong and fluctuating background and from exaggerated positional uncertainty for particles of interest being tracked at high speed with small probes. Here, the authors describe image-processing methodology that extracts the full point spread function for a nanoparticle under study from each individual image frame to increase positional precision. The methodology also allows the axial position of a nanoparticle to be calculated precisely for the purpose of 3D imaging. The authors label epidermal growth factor receptors with gold nanoparticles in HeLa cells and are able to track them with low nanometer spatial precision and midmicrosecond temporal resolution (at 20,000 frames/s) for periods of 10 min. These capabilities enabled description of the motions of EGFR as it diffuses within the plasma membrane, gets transported along filopodia, and becomes confined in clathrin-coated pits. The results reveal previously unstudied phenomena of molecular motions in biologic membranes.

**Chang B-J, Kittisopikul M, Dean KM, Roudot P, Welf ES, Fiolka R. Universal light-sheet generation with field synthesis. Nat Methods 2019;16:235-238.**

Light-sheet fluorescence microscopy has gained greatly in popularity as a methodology for high-resolution, high-speed imaging of live organisms and tissues over long periods of time. It works by illumination of the sample with a thin sheet of light in the imaging plane of the microscope, thereby minimizing background and radiation dosage to nonimaged portions of the sample. The light sheet is scanned across the sample during image acquisition. The method therefore constitutes an optical sectioning technique. However, there are tradeoffs among axial resolution, field of view, illumination duty cycle, and illumination confinement. Various illumination modalities, such as lattice light-sheet microscopy, have been developed to optimize these parameters simultaneously. These modalities have realized significant improvement in performance but at the expense of increased complexity of the optical train and increased cost. Chang *et al.* here describe a mathematical theorem they call "field synthesis," the application of which enables the generation of light sheets for such improved methods with much simpler optics. It creates desired light-sheet intensity profiles by summing light sheets through time averaging. This obviates the need for complex optics, enhances illumination efficiency, and achieves minimal rates of photobleaching. It also enables simultaneous multicolor illumination. It is hoped that the methodology will accelerate the adoption of light-sheet fluorescence microscopy in biology.

**Kishi JY, Lapan SW, Beliveau BJ, et al. SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. Nat Methods 2019;16:533-544.**

Kishi *et al.* seek to improve the utility of fluorescence *in situ* hybridization (FISH) for detection and quantification of RNA and DNA molecules in fixed tissues. In the present paper, they contribute methods by which to increase signal strength by amplification of single-stranded oligonucleotide FISH probes using a strand-displacing polymerase to form single-stranded DNA concatemers. Signal is produced by hybridization of fluorescently labeled oligonucleotides complementary to the concatemer sequences. Signal amplification is 5–450-fold in fixed cells and tissues. Specificity is high enough to allow simultaneous amplification of up to 17 different targets and efficient enough to detect mRNAs. Signal amplification enables testing of thick tissue slices, which generate high levels of autofluorescence, light scattering, and optical aberration that diminish signal-to-background ratios. Amplification also shortens imaging time.